Soc 361 Week 1: Introduction

LEARNING ABOUT THE SOCIAL WORLD USING STATISTICS

Topics:

I. Studying society: empirical vs. normative questions
II. Testing initial hypotheses about relationships among variables using data
III. Operationalization: concepts vs. measures of concepts
IV. Inference: Using statistics to generalize about a population based on a sample
V. Association vs. Causation vs. Determination
VI. Level of Measurement (type of variable)
VII. More Measurement Issues
VIII. Dependent and Independent Variables
IX. Measures of Central Tendency
X. Measures of Dispersion

I.Studying society: empirical vs. normative questions.

- An empirical question asks about how things are.
- A normative question asks about how things should be.
- How are normative and empirical questions related? Very debatable —

IMHO, circular — must have beliefs to interpret world, but normative must be empirical

II. *Testing* initial *hypotheses* about *relationships* among *variables* using *data*

- *Data*: Systematically gathered empirical information.
- *Variables*: Any concept that can take on two or more values or categories.
- *Relationships*: certain categories of one variable are associated with certain categories of another variable.
- *Hypothesis*: A statement which identifies at least two variables that are related and indicates how they are related.
- *Testing*, not proving: Are the data consistent with our hypotheses? Key, if obvious– we find evidence, not proof

III. Operationalization: concepts vs. measures of concepts.

- What is "income?"
- What are some possible measures of income? Think hrly, weekly, annual, pre-/post-txfer, capital vs. labor, inflation adj., PPP-adj.
- What are "views regarding gender roles?" Possible measures?
- What is "level of education"? Possible measures? – complex!

PROMPT: open GSS, look at
- educ
&
- degree –

IV. *Inference* : Using statistics to generalize about a *population* based on a *sample*

- Population: the entire set of persons, events, or other units that one wishes to make a statement about.
- Sample: any subset of a population.
- Most useful for inferences is a *random sample*, i.e. one in which every member of the population has an equal probability of being included.
- The members of the population define the *unit of analysis* of our study.

NB that most, ~ all, surveys are complex
- Stratif. & clustered
- random is useful theoretical simplif.

not always ppl!

Prompt: cases in GSS, auto, covid

IV. *Inference* : Using statistics to generalize about a *population* based on a *sample*

- Typically, we use the sample to obtain an *estimate* of some characteristic of the population.
  - Example: we use a survey sample to estimate the mean height, weight, education level, income, proportion female, etc. of a population.

"Population statistic" usually called a parameter → generically θ

$\mu_y : \bar{y}$; $\sigma_y : s_y$; $\rho_{x,y} : r_{x,y}$; $\pi : p$; $\beta : \hat{\beta}$ etc.

IV. *Inference*: Using statistics to generalize about a *population* based on a *sample*

- We use *inferential statistics* to make statements about the range in which a population characteristic ("parameter") probably falls, given the sample estimate.
  - Example: given that the sample mean on the variable y is x, the true population mean probably falls in the interval x-c to x+c.

more commonly, $\bar{y}$

these are based on std. dev. of sampling distro

IV. *Inference* : Using statistics to generalize about a *population* based on a *sample*

- We also obtain estimates and make inferences about relationships between variables in a population using a sample.
  - Example: given that the sample difference between group A and group B with respect to y equals x in the sample, the true population difference probably falls in the interval x-c to x+c.

V. Association vs. Causation vs. Determination

Prompt: tab marital role; coll age childs

- Association: variable x and variable y are related in some systematic way
- Causation: a change in variable x usually leads variable y to change in some systematic way
- Determination: a change in variable x inevitably produces a particular, predictable change in variable y.

Can we ever do better than assoc.?

yes, but rarely — see SOC362

VI. Level of Measurement

- What type of variable are we dealing with?
  - A. Nominal (discrete, categorical, qualitative).
  - B. Ordinal (ordered).
  - C. Interval (quantitative, continuous).

↳ Subtype of ratio: meaningful zero
— best for "pure quant." techniques

VI. Level of Measurement

usually not a prob. but use caution

- A. Nominal (discrete, categorical, qualitative) variables.

Prompt: d race; lab li RACE; graph bar, over(race)

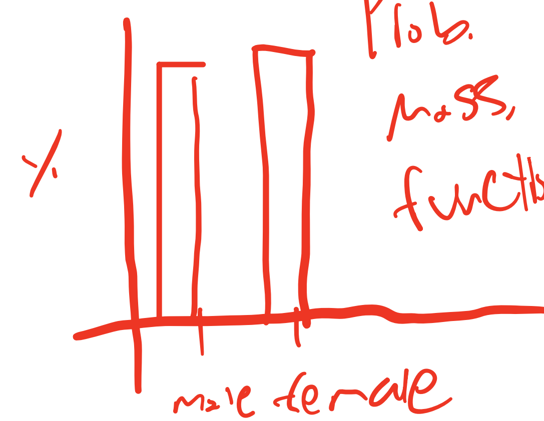
  - Categories represent different *attributes*, not *quantities*.
  - Numerical values assigned to categories are arbitrary.

but in many data sets use them anyway

  - Categories must be mutually exclusive and exhaustive.
  - Frequency distributions: grouped nominal data.
  - Examples? relig., race, sex, marital status
  - Special Case: Dichotomous Variables.

This race/gender/relig or not, etc.

Prob. mass function



- Dichotomous variables where the assigned values are zero and one are called “dummy” variables.

Also: binary, Boolean, Bernoulli... indicator

VI. Level of Measurement

- B. Ordinal (ordered) variables.

Prompt: tab polviews; graph bar, over(polv)

  - Categories are *intrinsically* ordered from lowest to highest, but they do not represent precise quantities.
  - Numerical values represent the order of the categories, but not the gap or distance between them, which is indeterminate.
  - Typical example: "Likert" scales.
    1.strongly agree
    2.agree
    3.disagree
    4.strongly disagree
  - Social class categories
    1. Upper class
    2. Middle class
    3. Working class
    4. Lower class
  - Other Examples?

many recoded quant. vars
e.g. relig. attendance,
age bracket, income bkt
many surveys do this

VI. Level of Measurement

Prompt: sum edu, detail; sum height, d; hist height if sex==1, kdens normal percent

- C. Interval (quantitative, continuous) variables.
  - Categories correspond to precise numerical scores.
  - Numerical values represent the exact score on the variable, distances between categories can be determined by subtraction.
  - "Ratio" variables are interval variables with a lower limit of zero.
  - Examples? interval: Celsius, years since left school;

ratio: weight, height, income

One other dist: discrete (limited range) vs. contin. (any real number $\in \mathbb{R}$) — all data are discrete, but parent var. can be cont.

VI. Level of Measurement

- D. Some general considerations regarding level of measurement.

Cf above

  - 1. Sometimes interval variables can be presented as grouped interval data.
  - 2. Usually interval variables are better summarized using measures of central tendency and dispersion (mean and standard deviation).

Do NOT MAKE THIS MISTAKE

  - 3. Measures of central tendency and dispersion have no meaning when the variables are not interval.

Prompt: Sum relig

  - 4. Some variables can be construed as either interval or ordinal, depending on the context. (Examples?)

+ one common debate: Likert

what about educ? is a year always a year?

VII. More Measurement Issues

A. Choosing the appropriate form of a variable.

B. Sources of *measurement error* in survey research.

- Poorly designed questions/questionnaires.
- Respondent error.
- Recording error.
- Data entry error.
- Vague concepts.

must think hard about this but not really covered at length in 361

VII. More Measurement Issues

C.Scales and indices.

These terms are rarely used consistently – more later

Example: Liberal vs. Conservative attitudes scale

1.Abortion should be illegal.
2.“Family Values” should be taught in schools.
3.Full funding for the Defense Department is needed for national security.
4.Education and welfare should be handled by the states, not the federal government.
5.It is wrong for the government to control or outlaw guns.

VIII. Dependent and Independent Variables

A. When we hypothesize a causal relationship among variables, we call the variable which is "caused by" or "depends on" some other variable or variables the *dependent* variable.

Also: response, outcome

- The dependent variable answers the question: what is it about different groups that is being compared?
- Consider the hypothesis: "whites are more likely to commit white collar crimes than non-whites." What is the dependent variable?

* probability of committing crime – not directly observed individ. – infer from mean of subgroup

VIII. Dependent and Independent Variables

Also: predictor, regressor

B. The variable or variables which we think influence the dependent variable are called *independent* variables. The independent variable defines the groups that are being compared.

What is the independent variable in the previous statement?

C. Another example: sex and age at marriage. Hypothesis: "women are likely to get married at younger ages than men."

- 1.What is the independent variable? Sex
- 2.What is the dependent variable? Age at (first?) marriage
- 3.How might we operationalize the dependent variable? Prob. year
- 4.If the hypothesis is true, what would we expect to find in the data?

$age_f - age_m < 0$ → and Stat. Sig. — more later

IX. Central Tendency

- A measure of central tendency describes a “most typical” or “most usual” score in distribution of scores on a variable.
- Purpose: To describe some quantitative characteristic of a group using a single, precise number.
- Examples:
  - Arithmetic mean
  - Median
  - Mode
  - (Geometric mean) $\sqrt[n]{\prod_{i=1}^{n} y_i} = (y_1 * y_2 * y_3 \ldots y_{n-1} * y_n)^{1/n}$

The arithmetic mean

- The mean (average) is the most common measure of central tendency.
- The mean is computed by summing up the values on a variable, *x*, for all the observations and dividing by the sample size (n).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

The arithmetic mean

sometimes $\hat{\mu}_x$ ≈ "input"

- In this formula:
  - The "x-bar" refers to the mean
  - The Σ is the "summation" sign – read it as saying "the sum of all [argument] where [argument] refers to whatever follows the summation sign.
  - The i subscript on the x in the [argument] denotes the value of x on observation number i. So, I can range from 1 (the first observation) to n (the nth) observation). That is what the "i=1" under the summation and the n on top of the summation mean.
  - This formula may look complicated at first, but it is actually must simpler than writing the full formula for the mean (on the right).

almost always for us, start at 1, but sometimes 0

can omit lower & upper indices if clear

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

$\sum_{i=1}^{n} K$

summ. rules

1. $\sum_{i=1}^{n} K y_i = K \sum y_i$

2. $\sum (y_i + z_i) = \sum y_i + \sum z_i$

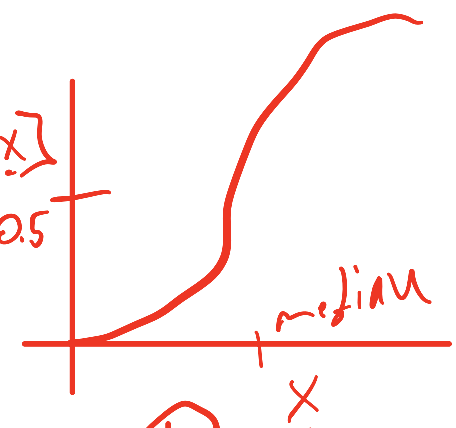
3. $= nK$

The median

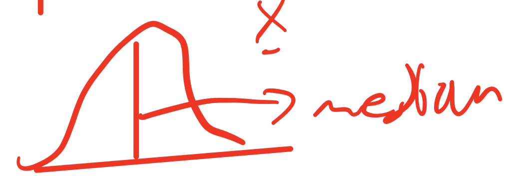
- The median:
  - The value of x for which there are an equal number of scores above and below the value in the distribution of x.
  - The "middle" score in a distribution.
  - The 50^{th} percentile.

On a cumulative dist function, $X : F(x) = 0.5$ $[P[X \le x] = 0.5]$

Normal CDF



on a prob. dens. function, equal areas pt.



The median

- To compute the median:
  - First, rank all the observations in a distribution in order from lowest to highest values on the variable whose median you wish to determine.
  - Then, divide the sample size in half.
  - Then, count that many observations from the beginning of the ranked list.
  - If n is odd, then the median is the midway point between the two observations that fall on either side of the halfway point.

Other software uses other interpol. formula

"Valid percent" is an archaic SPSS thing

The median

- Using a frequency table:
  - The median is the value on x for which the cumulative percent equals 50.
  - If there is no cumulative percent exactly equal to 50, use the first cumulative percent that is greater than 50.
  - For example, the median years of education in the data to the right is 12.0.

EDUC

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	105	.3	.3	.3
	1	31	.1	.1	.4
	2	79	.2	.2	.6
	3	199	.5	.5	1.1
	4	255	.7	.7	1.8
	5	327	.9	.9	2.6
	6	533	1.4	1.4	4.0
	7	728	1.9	1.9	5.9
	8	2170	5.7	5.7	11.7
	9	1450	3.8	3.8	15.5
	10	2033	5.3	5.4	20.8
	11	2404	6.3	6.3	27.1
	12	12235	32.1	32.2	59.3
	13	2993	7.9	7.9	67.2
	14	3643	9.6	9.6	76.8
	15	1541	4.0	4.1	80.9
	16	4118	10.8	10.8	91.7
	17	1026	2.7	2.7	94.4
	18	1072	2.8	2.8	97.2
	19	437	1.1	1.2	98.4
	20	619	1.6	1.6	100.0
	Total	37998	99.7	100.0	
Missing	98 DK	57	.1		
	99 NA	61	.2		
	Total	118	.3		
Total		38116	100.0		

The mode

- The mode is simply the value of x in which largest proportion of cases fall.
- Just look at the frequency distribution to identify the mode.
- The mode need not have a majority of cases.

The mode

- What is the mode for this distribution? educ== 12
- What percentage of cases have the modal value? 0.322

EDUC

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	105	.3	.3	.3
	1	31	.1	.1	.4
	2	79	.2	.2	.6
	3	199	.5	.5	1.1
	4	255	.7	.7	1.8
	5	327	.9	.9	2.6
	6	533	1.4	1.4	4.0
	7	728	1.9	1.9	5.9
	8	2170	5.7	5.7	11.7
	9	1450	3.8	3.8	15.5
	10	2033	5.3	5.4	20.8
	11	2404	6.3	6.3	27.1
	12	12235	32.1	32.2	59.3
	13	2993	7.9	7.9	67.2
	14	3643	9.6	9.6	76.8
	15	1541	4.0	4.1	80.9
	16	4118	10.8	10.8	91.7
	17	1026	2.7	2.7	94.4
	18	1072	2.8	2.8	97.2
	19	437	1.1	1.2	98.4
	20	619	1.6	1.6	100.0
	Total	37998	99.7	100.0	
Missing	98 DK	57	.1		
	99 NA	61	.2		
	Total	118	.3		
Total		38116	100.0		

Central tendencies: some general considerations

- Remember: only describe *interval* variables with the mean. The median may be used for interval or ordinal variables. Only the mode can be used to describe nominal variables.

- Remember: the point of computing the mean or median is ultimately to compare one group to other groups.

mean is natural: balance point of PDF;

- In general, the mean is the preferred measure of central tendency for interval variables. But when the distribution is highly *skewed* (i.e., the distribution is asymmetric and the mean is much higher or much lower than the median), then the median is preferable.

least squares estimation

$$\frac{d \sum_{i}^{n} (y_i - \theta)^2}{d\theta} = -2\sum_i (y_i - \theta);$$

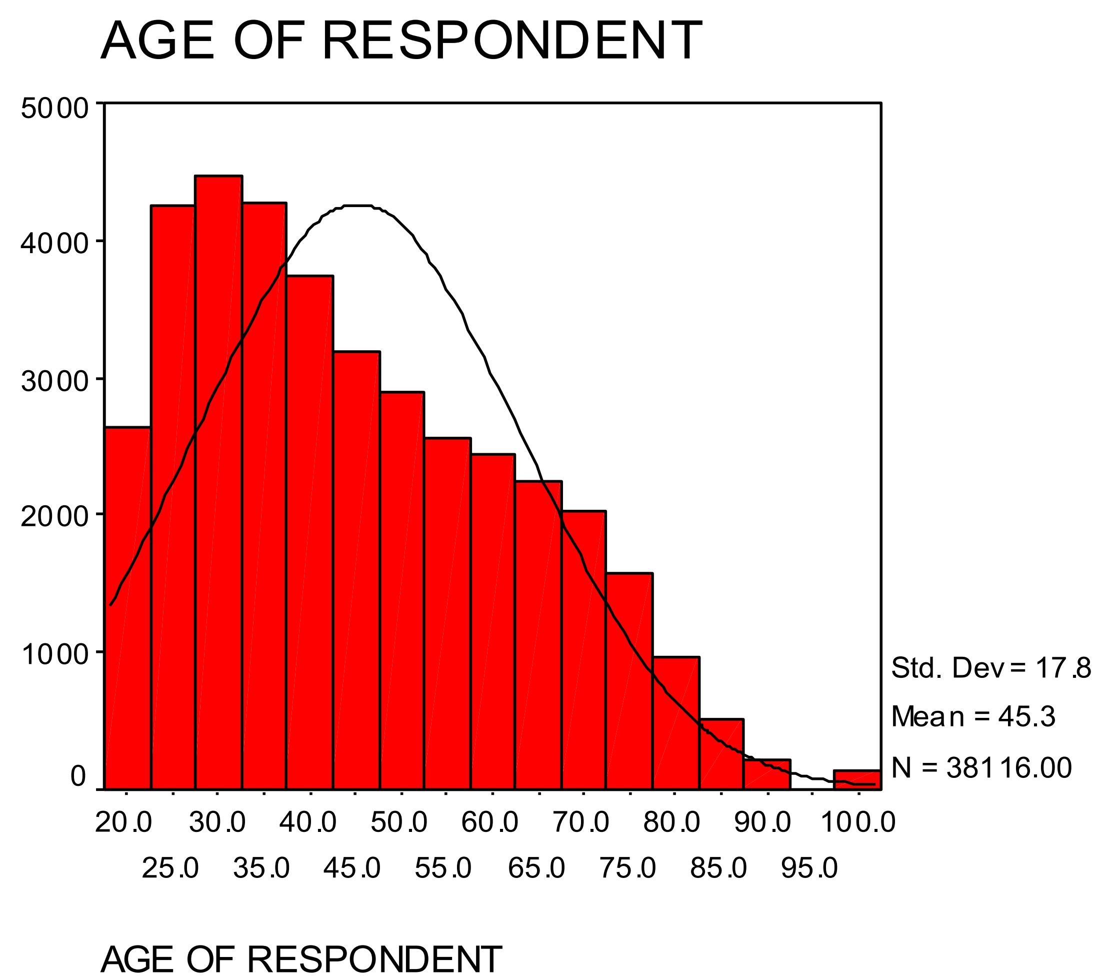
set to zero, & $\sum y_i = n\theta$; $\theta = \bar{y}$

Central tendencies: some general considerations

Prompt: use cps, sum wage1

- Here is a positively skewed distribution:

AKA right skew

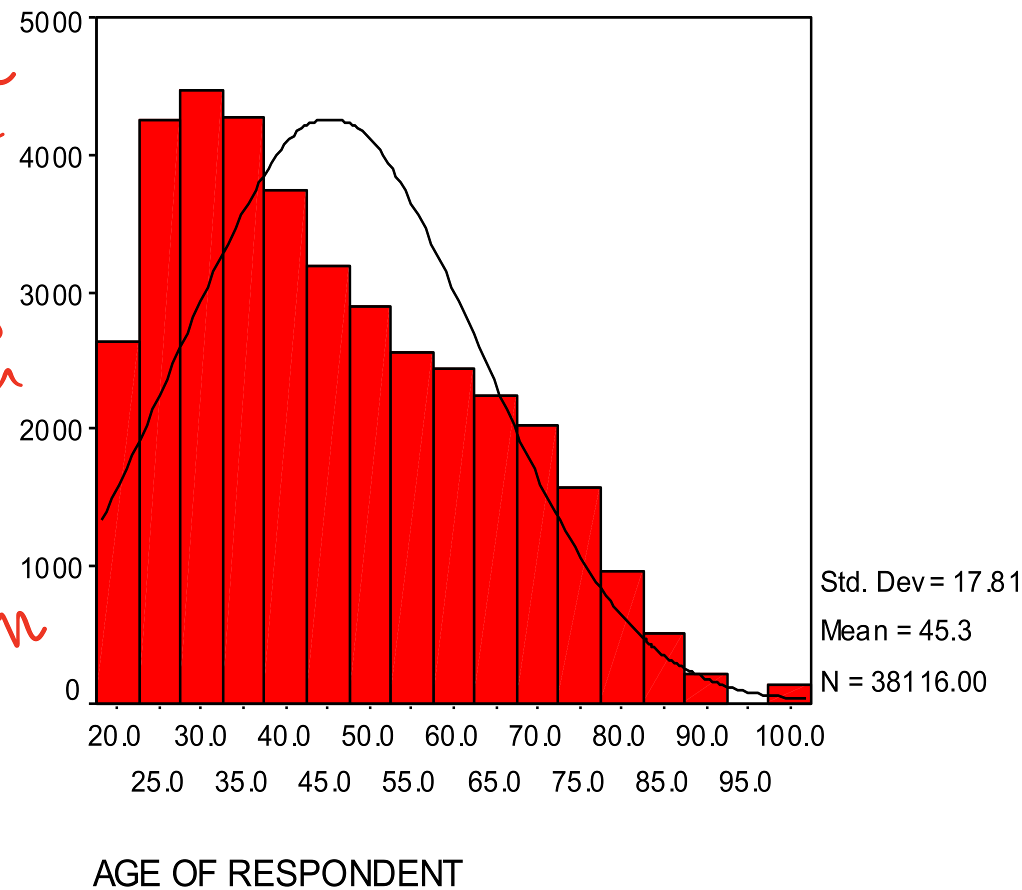


Central tendencies: some general considerations

- Note that the mean is pulled upward by the extremely high values.
- The median will be smaller than mean. It is 42 for the distribution to the right.
- Because the median is not affected by extreme values (why not?) it is usually preferable when dealing with a skewed distribution.
- Examples of distributions that are usually skewed:
  - Income
  - Hospital visits
  - Years spent in prison

add one extreme value & it does not even shift up a full person

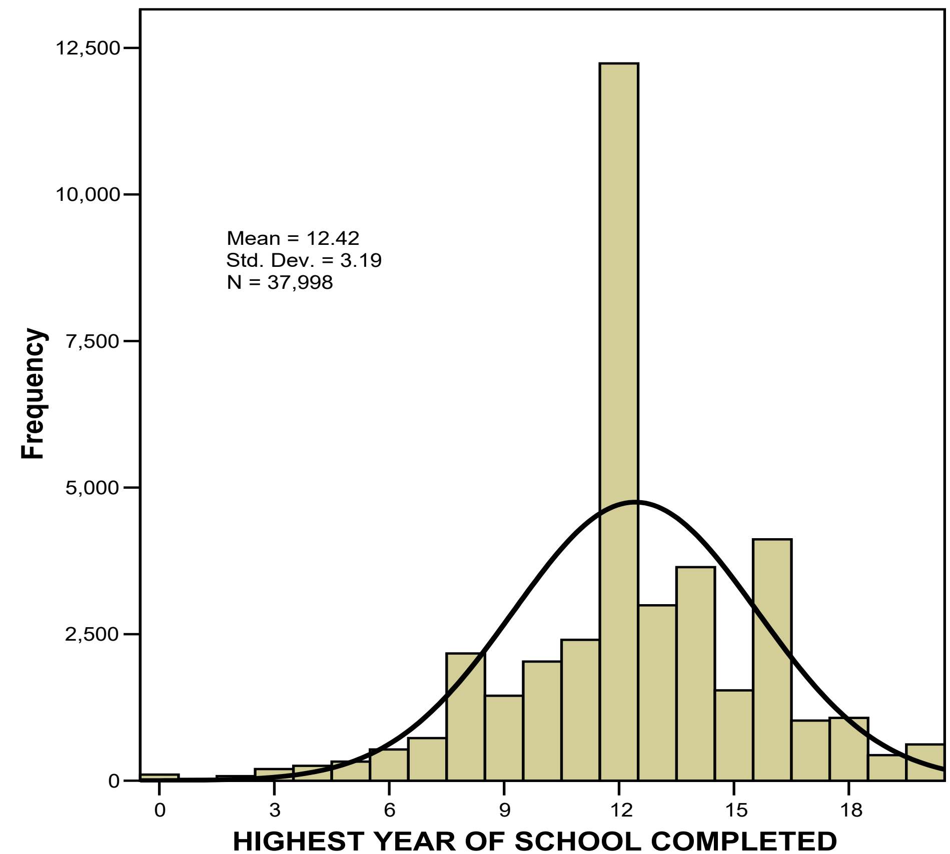
AGE OF RESPONDENT



Central tendencies: some general considerations

educ   HIGHEST YEAR OF SCHOOL COMPLETED

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	105	.3	.3	.3
	1	31	.1	.1	.4
	2	79	.2	.2	.6
	3	199	.5	.5	1.1
	4	255	.7	.7	1.8
	5	327	.9	.9	2.6
	6	533	1.4	1.4	4.0
	7	728	1.9	1.9	5.9
	8	2170	5.7	5.7	11.7
	9	1450	3.8	3.8	15.5
	10	2033	5.3	5.4	20.8
	11	2404	6.3	6.3	27.1
	12	12235	32.1	32.2	59.3
	13	2993	7.9	7.9	67.2
	14	3643	9.6	9.6	76.8
	15	1541	4.0	4.1	80.9
	16	4118	10.8	10.8	91.7
	17	1026	2.7	2.7	94.4
	18	1072	2.8	2.8	97.2
	19	437	1.1	1.2	98.4
	20	619	1.6	1.6	100.0
	Total	37998	99.7	100.0	
Missing	98 DK	57	.1		
	99 NA	61	.2		
	Total	118	.3		
Total		38116	100.0		



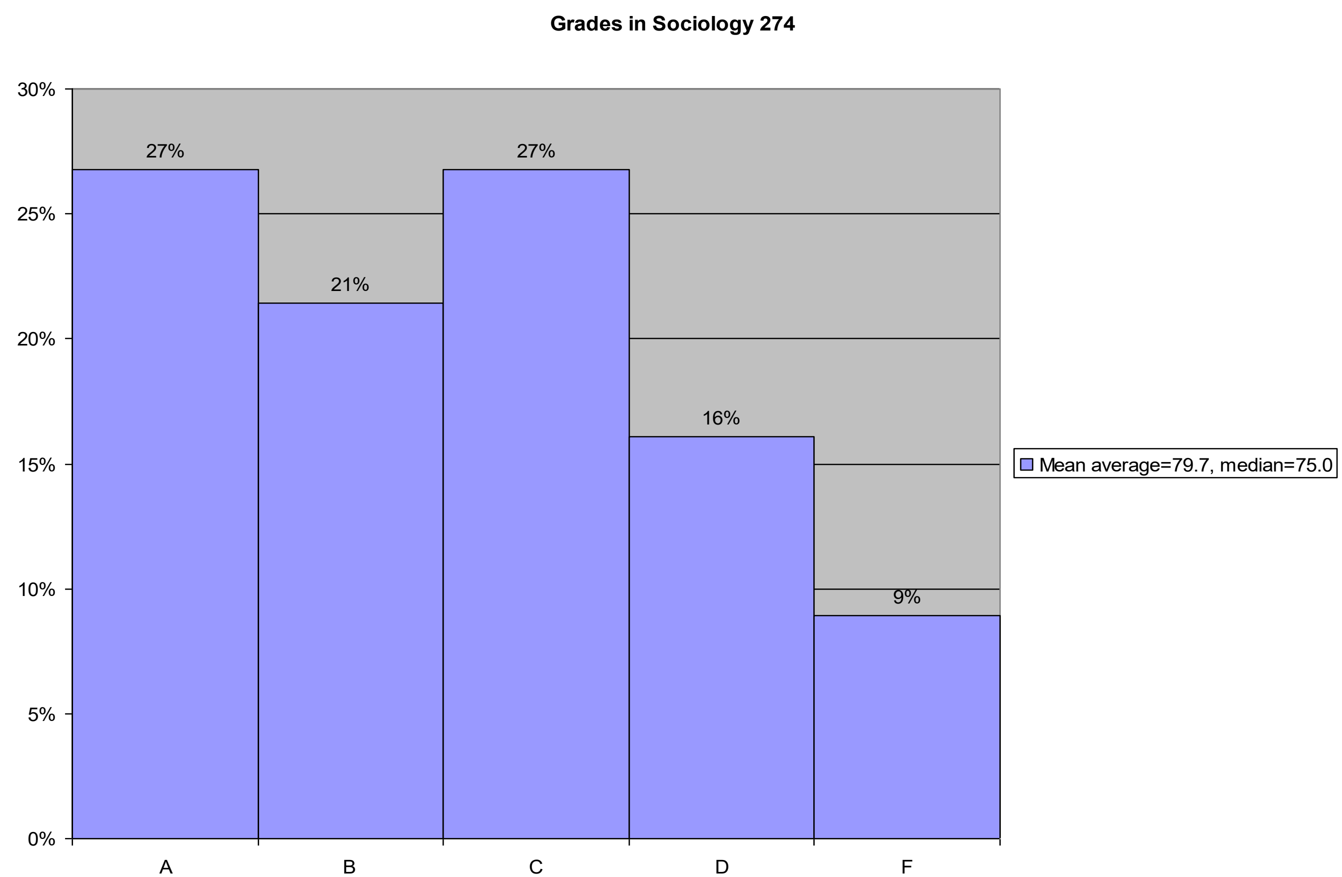
AKA left

- Here is a variable with (slight) negative skew.
12
  - What is the median? Is it lower/higher than the mean?

lower - ine doesn't

Bimodal distribution

- Here is another type of distribution, a “bi-modal” distribution:



X.Measures of Dispersion

- **Purpose: to describe how clustered or scattered about the mean a group's scores on a variable are (how dispersed is the distribution).**
- **Examples:**
  - **The range**
  - **The inter-quartile range**
  - **Variance**
  - **Standard deviation**

The Range

- The distance between the maximum and minimum values of the variable for the group is called the **range.**
- To determine the range, determine the lowest and highest values by ranking the cases, adjust for rounding, and subtract the lowest from the highest.

Not used much

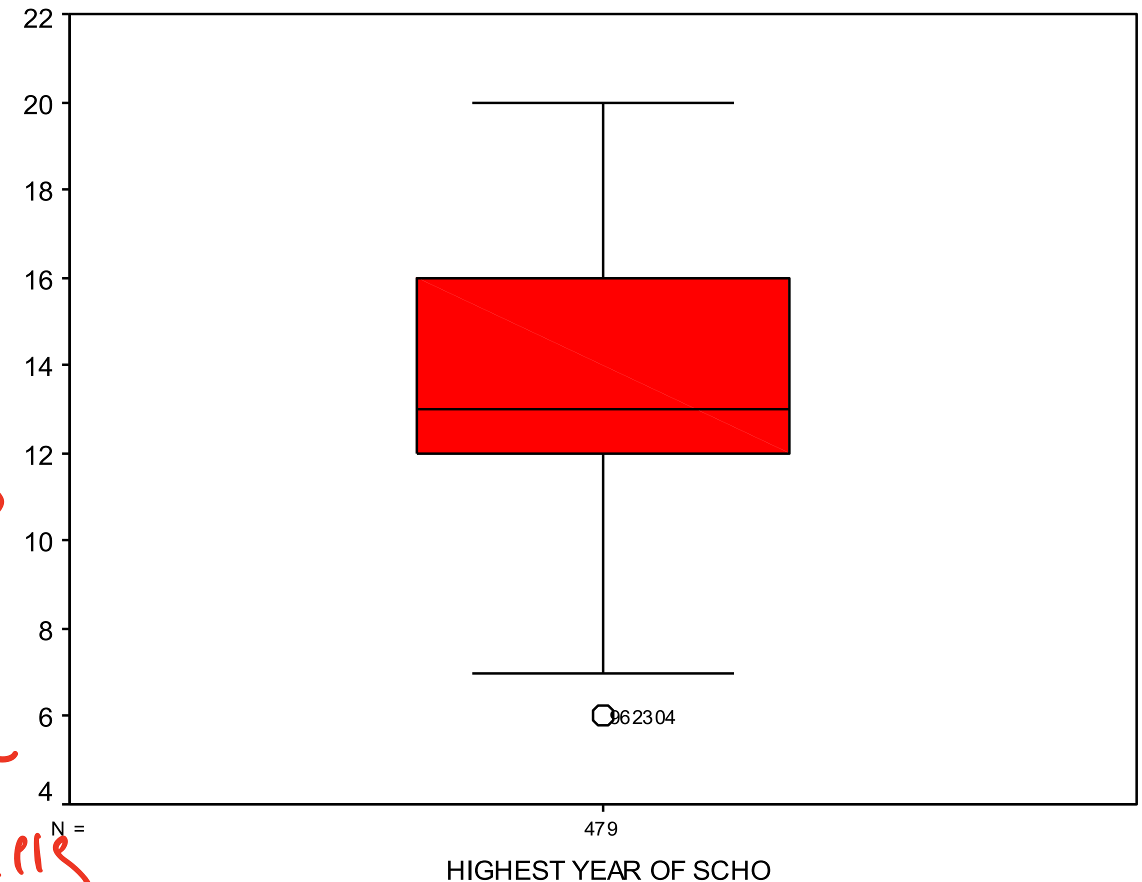
Inter-quartile range

- The distance between the 25^{th} and 75^{th} percentiles is called the “inter-quartile range.”
  - This is more informative than the range, because the range is easily affected by an extremely high or extremely low value.
  - The inter-quartile range tells you how dispersed about the median are those 50% of cases that are closest to the median.

A bit — one outlier

Inter-quartile range

- Remember that the inter-quartile range is displayed in a box plot.
- Does this variable look skewed? Explain.



Stand. box plots use Tukey rule for outliers: if $y_i > Q_3 + 1.5(IQR)$ $\cup$ $y_i < Q_1 - 1.5(IQR)$

Prompt: Show tukey rule

Variance and standard deviation

- The most common and useful measures of dispersion are the variance and standard deviation, which are closely related:

  - Variance = Standard deviation squared = s^2

  - Standard deviation = Square-root of the variance $= \sqrt{\text{var}}$

Variance – easier to use in proofs; std dev more interp.

Variance and standard deviation: formulas for samples

$$Variance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

why square

this is the length of our data considered as a centered vector in n-dim space (divided by n-1)

Also: Sum of devs. alone is 0!

$$St.Dev. = \sqrt{Variance} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Proof: $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum x_i - \sum \bar{x}$

How to compute a variance/standard deviation

$= n\bar{x} - n\bar{x}$

$= 0$

- Compute the variance and standard deviation for samples 1 and 2.

SAMPLE 1					SAMPLE 2			
	Years of					Years of		
	Education					Education		
	x					x		
x_1	10				x_1	11		
x_2	12				x_2	12		
x_3	16				x_3	13		
x_4	16				x_4	13		
x_5	12				x_5	14		
x_6	16				x_6	16		
x_7	10				x_7	13		
x_8	14				x_8	12		
x_9	14				x_9	12		
x_{10}	8				x_{10}	12		

How to compute a variance/standard deviation

- 1.Compute the mean by summing up the scores and dividing by n.

SAMPLE 1					SAMPLE 2			
	Years of					Years of		
	Education					Education		
	x					x		
x_1	10				x_1	11		
x_2	12				x_2	12		
x_3	16				x_3	13		
x_4	16				x_4	13		
x_5	12				x_5	14		
x_6	16				x_6	16		
x_7	10				x_7	13		
x_8	14				x_8	12		
x_9	14				x_9	12		
x_{10}	8				x_{10}	12		
Σ	128				Σ	128		
$\Sigma/(n)$	12.8				$\Sigma/(n)$	12.8		

How to compute a variance/standard deviation

- 2. Compute the deviation of each score from the mean by subtracting the mean.

I told you!

SAMPLE 1					SAMPLE 2			
	Years of					Years of		
	Education					Education		
	x	x-mean(x)				x	x-mean(x)	
x_1	10	-2.8			x_1	11	-1.8	
x_2	12	-0.8			x_2	12	-0.8	
x_3	16	3.2			x_3	13	0.2	
x_4	16	3.2			x_4	13	0.2	
x_5	12	-0.8			x_5	14	1.2	
x_6	16	3.2			x_6	16	3.2	
x_7	10	-2.8			x_7	13	0.2	
x_8	14	1.2			x_8	12	-0.8	
x_9	14	1.2			x_9	12	-0.8	
x_{10}	8	-4.8			x_{10}	12	-0.8	
Σ	128	0			Σ	128	0	
Σ/(n)	12.8				Σ/(n)	12.8		

How to compute a variance/standard deviation

- 3. Square each of the deviations.

* Sum of Sq.

— We'll use often!

SAMPLE 1					SAMPLE 2			
	Years of Education					Years of Education		
	x	x-mean(x)	$(x\text{-mean}(x))^2$			x	x-mean(x)	$(x\text{-mean}(x))^2$
x_1	10	-2.8	7.84		x_1	11	-1.8	3.24
x_2	12	-0.8	0.64		x_2	12	-0.8	0.64
x_3	16	3.2	10.24		x_3	13	0.2	0.04
x_4	16	3.2	10.24		x_4	13	0.2	0.04
x_5	12	-0.8	0.64		x_5	14	1.2	1.44
x_6	16	3.2	10.24		x_6	16	3.2	10.24
x_7	10	-2.8	7.84		x_7	13	0.2	0.04
x_8	14	1.2	1.44		x_8	12	-0.8	0.64
x_9	14	1.2	1.44		x_9	12	-0.8	0.64
x_{10}	8	-4.8	23.04		x_{10}	12	-0.8	0.64
Σ	128	0	73.60		Σ	128	0	17.60
Σ/(n)	12.8				Σ/(n)	12.8		

How to compute a variance/standard deviation

- 4. Divide the sum of squared deviations by n-1 for the variance.

SAMPLE 1					SAMPLE 2			
	Years of Education					Years of Education		
	x	x-mean(x)	$(x\text{-mean}(x))^2$			x	x-mean(x)	$(x\text{-mean}(x))^2$
x_1	10	-2.8	7.84		x_1	11	-1.8	3.24
x_2	12	-0.8	0.64		x_2	12	-0.8	0.64
x_3	16	3.2	10.24		x_3	13	0.2	0.04
x_4	16	3.2	10.24		x_4	13	0.2	0.04
x_5	12	-0.8	0.64		x_5	14	1.2	1.44
x_6	16	3.2	10.24		x_6	16	3.2	10.24
x_7	10	-2.8	7.84		x_7	13	0.2	0.04
x_8	14	1.2	1.44		x_8	12	-0.8	0.64
x_9	14	1.2	1.44		x_9	12	-0.8	0.64
x_{10}	8	-4.8	23.04		x_{10}	12	-0.8	0.64
Σ	128	0	73.60		Σ	128	0	17.60
Σ/(n)	12.8				Σ/(n)	12.8		
Σ/(n-1)			8.18		Σ/(n-1)			1.96

How to compute a variance/standard deviation

- 5. Take the square root of the variance for the standard deviation.

SAMPLE 1					SAMPLE 2			
	Years of Education					Years of Education		
	x	x-mean(x)	$(x\text{-mean}(x))^2$			x	x-mean(x)	$(x\text{-mean}(x))^2$
x_1	10	-2.8	7.84		x_1	11	-1.8	3.24
x_2	12	-0.8	0.64		x_2	12	-0.8	0.64
x_3	16	3.2	10.24		x_3	13	0.2	0.04
x_4	16	3.2	10.24		x_4	13	0.2	0.04
x_5	12	-0.8	0.64		x_5	14	1.2	1.44
x_6	16	3.2	10.24		x_6	16	3.2	10.24
x_7	10	-2.8	7.84		x_7	13	0.2	0.04
x_8	14	1.2	1.44		x_8	12	-0.8	0.64
x_9	14	1.2	1.44		x_9	12	-0.8	0.64
x_{10}	8	-4.8	23.04		x_{10}	12	-0.8	0.64
Σ	128	0	73.60		Σ	128	0	17.60
Σ/(n)	12.8				Σ/(n)	12.8		
Σ/(n-1)			8.18		Σ/(n-1)			1.96
SQRT(Σ/(n-1))			2.86		SQRT(Σ/(n-1))			1.40

How to compute a variance/standard deviation

- How are the samples similar? How do they differ?

SAMPLE 1					SAMPLE 2			
	Years of Education					Years of Education		
	x	x-mean(x)	$(x\text{-mean}(x))^2$			x	x-mean(x)	$(x\text{-mean}(x))^2$
x_1	10	-2.8	7.84		x_1	11	-1.8	3.24
x_2	12	-0.8	0.64		x_2	12	-0.8	0.64
x_3	16	3.2	10.24		x_3	13	0.2	0.04
x_4	16	3.2	10.24		x_4	13	0.2	0.04
x_5	12	-0.8	0.64		x_5	14	1.2	1.44
x_6	16	3.2	10.24		x_6	16	3.2	10.24
x_7	10	-2.8	7.84		x_7	13	0.2	0.04
x_8	14	1.2	1.44		x_8	12	-0.8	0.64
x_9	14	1.2	1.44		x_9	12	-0.8	0.64
x_{10}	8	-4.8	23.04		x_{10}	12	-0.8	0.64
Σ	128	0	73.60		Σ	128	0	17.60
Σ/(n)	12.8				Σ/(n)	12.8		
Σ/(n-1)			8.18		Σ/(n-1)			1.96
SQRT(Σ/(n-1))			2.86		SQRT(Σ/(n-1))			1.40

Same mean; much higher var. in 1st one

How to compute a variance/standard deviation

Key

- NOTE!!: for a population, you divide the sum of squared deviations by n, rather than n-1.

GROUP 1 (whole population)					GROUP 2 (whole population)			
	Years of Education					Years of Education		
	x	x-mean(x)	$(x\text{-mean}(x))^2$			x	x-mean(x)	$(x\text{-mean}(x))^2$
x_1	10	-2.8	7.84		x_1	11	-1.8	3.24
x_2	12	-0.8	0.64		x_2	12	-0.8	0.64
x_3	16	3.2	10.24		x_3	13	0.2	0.04
x_4	16	3.2	10.24		x_4	13	0.2	0.04
x_5	12	-0.8	0.64		x_5	14	1.2	1.44
x_6	16	3.2	10.24		x_6	16	3.2	10.24
x_7	10	-2.8	7.84		x_7	13	0.2	0.04
x_8	14	1.2	1.44		x_8	12	-0.8	0.64
x_9	14	1.2	1.44		x_9	12	-0.8	0.64
x_{10}	8	-4.8	23.04		x_{10}	12	-0.8	0.64
Σ	128	0	73.60		Σ	128	0	17.60
Σ/(n)	12.8				Σ/(n)	12.8		
Σ/(n)			7.36		Σ/(n)			1.76
SQRT(Σ/n)			2.71		SQRT(Σ/n)			1.33

Prompt: let's go w/ auto

Population vs. sample

- Remember: when asked to compute a variance and/or standard deviation, you must first determine whether the data you have are from a sample or from a population.
  - For a sample: divide by n-1.
  - For a population: divide by n.

Means and Standard Deviations For Dichotomous Variables

- We can code a dichotomous variable so that one of the categories is assigned a 0 and the other is assigned a 1.
- This is called a “dummy” variables.
- We do this because dummy variable have some useful properties:
    - The mean of a dummy variable gives you the proportion of cases in the “1” category.
    - The variance of dummy variable is equal to the proportion of cases in the “1” category times the proportion in the “0” category.
- More formally:

Means and Standard Deviations For Dichotomous Variables

Prompt: show w/ auto

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n};$$

Proof, a bit more slowly

$\therefore n\bar{X} = \sum_i^{n_j} \sum_j^{1} X_{ij}$

if $j \in [0,1]$, $= \sum_i^{n_j} X_{i0} = 0$

$+ \sum_i^{n_j} X_{i1} = n\bar{X}$,

$$\sum_{i=1}^{n} x_i = number\ of\ cases\ which\ equal\ 1,$$

$$\therefore\ \bar{x} = proportion\ of\ cases\ which\ equal\ 1.$$

Means and Standard Deviations For Dichotomous Variables

Proof: Key var. identity in pop. — König-Huygens

$$s^2 = (p_0)*(p_1) = (p_1)*(1-p_1) = (p_0)*(1-p_0) = \bar{x}*(1-\bar{x})$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2 = \frac{1}{N}\left(\sum y_i^2 - 2\mu\sum y_i + N\mu^2\right)$$

$$= \mu_{y^2} - \frac{(2N\mu^2 + N\mu^2)}{N}$$

$$= \mu_{y^2} - \mu^2$$

this is actually only approx. for sample!

- In this formula, p_0 is the proportion of “zeros” and p_1 is the proportion of “ones.”
- The variance equals the proportion of “ones” times the proportion of “zeros.”

Means and Standard Deviations For Dichotomous Variables

a 0,1 var squared

$$s = \sqrt{(p_0)*(p_1)} = \sqrt{(p_1)*(1-p_1)} = \sqrt{(p_0)*(1-p_0)} = \sqrt{\bar{x}*(1-\bar{x})}$$

just returns itself!

So, we have $p - p^2 = p(1-p)$

- To get the standard deviation, just take the square root of the variance!

For sample, see below

Standardizing a variable

- We can *standardize* a variable by transforming it into a set of "*z-scores*."
- To do this, first subtract the mean from each score, then divide it by the standard deviation.

Show w/ auto

$$z_{x_i} = \frac{(x_i - \bar{x})}{s_x}$$

NB this does not require Normality!

Standardizing a variable

- This transforms each observation's raw score on x into units of standard deviations above or below the mean for the sample or group.

- For example:
  - A value of 0.0 on the standardized score means that the observation has a value equal to the mean on the original variable.
  - A value of +1.0 on the standardized score means the observation has a value one standard deviation above the mean on the original variable
  - A value of -2.3 on the standardized score means the observation has a value 2.3 standard deviations below the mean on the original variable

- The overall mean of the z scores for the sample will equal zero, and the standard deviation will equal 1.

Mean proof: each var is a mean-dev.; see above

SD proof: 1

$$\sigma_z^2 = \frac{1}{N}\sum\left\{\frac{(y_i-\mu_y)-0}{\sigma_y}\right\}^2$$

$$= \frac{1}{\sigma_y^2}\sum\frac{(y_i-\mu_y)^2}{N}$$

$$= \sigma_y^2/\sigma_y^2$$

$$= 1$$

Sample Var of a dummy:

$$S_y^2 = \frac{1}{n-1}\sum(y_i-\bar{y})^2$$

$$= \frac{1}{n-1}\left(\sum y_i^2 - 2\bar{y}\sum y_i + n\bar{y}^2\right)$$

$$= \frac{n}{n-1}\overline{y^2} - \frac{2n\bar{y}^2}{n-1} + \frac{n\bar{y}^2}{n-1}$$

$$= \frac{n}{n-1}\left(\overline{y^2} - \bar{y}^2\right)$$